PLOS GENETICS

METHODS

The Triangulation WIthin a STudy (TWIST) framework for causal inference within pharmacogenetic research

Jack Bowden[1]*, **Luke C Pilling**[2], **Deniz Türkmen**[2], **Chia-Ling Kuo**[3], **David Melzer**[2]

1 Exeter Diabetes Group (ExCEED), College of Medicine and Health, University of Exeter, Exeter, United Kingdom, **2** Epidemiology and Public Health Group, College of Medicine and Health, University of Exeter, Exeter, United Kingdom, **3** Connecticut Convergence Institute for Translation in Regenerative Engineering, University of Conneticut, Farmington, Conneticut, United States of America

* j.bowden2@exeter.ac.uk

Citation: Bowden J, Pilling LC, Türkmen D, Kuo C-L, Melzer D (2021) The Triangulation WIthin a STudy (TWIST) framework for causal inference within pharmacogenetic research. PLoS Genet 17(9): e1009783. https://doi.org/10.1371/journal.pgen.1009783

Editor: Zoltán Kutalik, University Hospital of the Canton Vaud (CHUV), SWITZERLAND

Received: April 28, 2021

Accepted: August 17, 2021

Published: September 8, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009783

Data Availability Statement: The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank (https://www.ukbiobank.ac.uk/). The derived data fields used in

Abstract

In this paper we review the methodological underpinnings of the general pharmacogenetic approach for uncovering genetically-driven treatment effect heterogeneity. This typically utilises only individuals who are treated and relies on fairly strong baseline assumptions to estimate what we term the 'genetically moderated treatment effect' (GMTE). When these assumptions are seriously violated, we show that a robust but less efficient estimate of the GMTE that incorporates information on the population of untreated individuals can instead be used. In cases of partial violation, we clarify when Mendelian randomization and a modified confounder adjustment method can also yield consistent estimates for the GMTE. A decision framework is then described to decide when a particular estimation strategy is most appropriate and how specific estimators can be combined to further improve efficiency. Triangulation of evidence from different data sources, each with their inherent biases and limitations, is becoming a well established principle for strengthening causal analysis. We call our framework '**T**riangulation **WI**thin a **ST**udy' (TWIST)' in order to emphasise that an analysis in this spirit is also possible within a single data set, using causal estimates that are approximately uncorrelated, but reliant on different sets of assumptions. We illustrate these approaches by re-analysing primary-care-linked UK Biobank data relating to *CYP2C19* genetic variants, Clopidogrel use and stroke risk, and data relating to *APOE* genetic variants, statin use and Coronary Artery Disease.

Author summary

Understanding how much a specific treatment's effect is moderated by common genetic variation is an important public health question. If a person's genetics means they will experience a much reduced treatment effect, as measured with respect to a particular health outcome, then they could be switched to an alternative therapy. When assessing the impact of such a switch at the population level, it is typical to only use data on those who are treated with the said drug. However, this analysis is compromised if genetic variants

our analysis will be available via the UK Biobank, searching for application number 14631 - we are not able to share these directly. An R package to implement TWIST is currently under development at https://github.com/lukepilling/twistR.

Funding: Jack Bowden is funded by an Expanding Excellence in England (E3) grant awarded to the University of Exeter. Chia-Lin Kuo, Luke C Pilling and David Melzer are supported in part by an R21 grant (R21AG060018) funded by National Institute on Aging, National Institute of Health, USA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

exist which moderate the treatment effect and affect the outcome through alternative pathways. In this paper we describe an extended analysis framework to estimating the 'genetically moderated treatment effect' (GMTE) that incorporates information on both treated and untreated individuals. With this larger set of information we show that four analysis approaches for estimating the GMTE are possible. Each one relies on a different set of assumptions to work correctly and provides estimates that are largely uncorrelated with one another. Our paper describes a decision framework for triangulating the findings from these four approaches in order to provide a more robust basis for decision making in public health.

1 Background

Over the last 20 years the field of Epidemiology has embraced the exploitation of random genetic inheritance to help uncover causal mechanisms of disease using the technique of Mendelian randomization (MR). [1]. The basic premise of MR is illustrated by the causal diagrams in Fig 1: Genetic variants, usually Single Nucleotide Polymorphisms or SNPs, G, are found which robustly explain variation in a modifiable risk factor, X, where X is typically continuous (for example a person's body mass index). The association between the exposure and an outcome, Y, hypothesised to be a downstream consequence of X, may be contributed to in observational data by unobserved confounding, U. If present, such confounding would mean that the naive association between X and Y would not reflect the causal effect of X on Y. If the important confounders could be appropriately measured and adjusted for, and no systematic selection bias or loss to follow up was present in the data this last assumption, then individuals with the same exposure level would be exchangeable [2] and observational associations could be interpreted causally. An MR analysis aims to circumvent any potential confounding by instead measuring the association between the outcome and the portion of the exposure that can be genetically predicted by the SNPs. Provided that the SNPs are independent of the confounders, are not associated with the outcome through any other pathway except the exposure, and the causal effect of a unit increase in the exposure is the same across individuals, then MR can consistently estimate the average causal effect of intervening on the exposure (e.g. to reduce or increase it) on the outcome. The Instrumental Variable (IV) assumptions of the MR approach are denoted IV1-IV4 in Fig 1.

Genetic variants can also play an important role in helping to explain treatment effect heterogeneity in pharmacogenetics research. A canonical example is Clopidogrel: the primary drug for ischemic stroke prevention in the UK and many other countries (see https://cks.nice.org.uk/antiplatelet-treatment). It requires CYP2C19 enzyme activation in order to be properly metabolised into the active form of the drug and thus work to its fullest extent. However, it has long been known that both common loss-of-function and gain-of-function variants within the *CYP2C19* gene region can massively impact each patient's ability to metabolise the drug [3]. Consequently, when prescribed in a primary care setting its effectiveness is heterogeneous, working well for some and not for others. Estimating the true effectiveness of a treatment from observational data is challenging due to 'confounding by indication' [4] (see Fig 1). For example, Clopidogrel use will, quite rightly, strongly depend on an individual's underlying risk of stroke. However, unmeasured socio-economic factors may also influence both an individual's ability to access appropriate healthcare and their underlying stroke risk [5]. Use of the drug for a sustained period may additionally depend on whether it can be tolerated without side-effects. Whilst these confounding factors could in principle be directly accounted for in a statistical

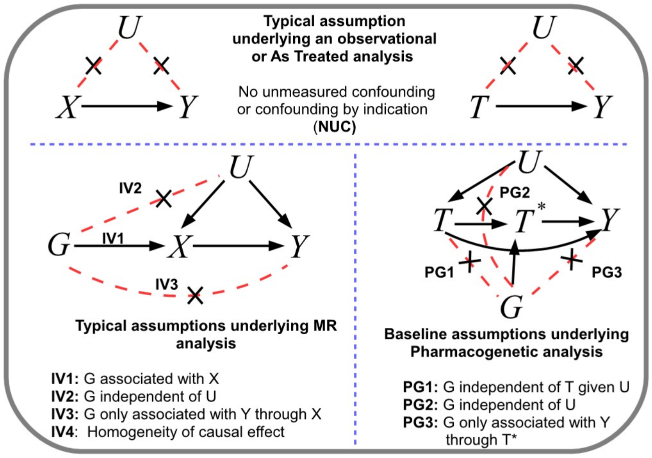
U
X
Y
Typical assumption underlying an observational or As Treated analysis
No unmeasured confounding or confounding by indication (NUC)
U
T
Y
U
IV2
G
IV1
X
Y
IV3
Typical assumptions underlying MR analysis
IV1: G associated with X
IV2: G independent of U
IV3: G only associated with Y through X
IV4: Homogeneity of causal effect
U
PG2
T
T*
Y
PG1
PG3
G
Baseline assumptions underlying Pharmacogenetic analysis
PG1: G independent of T given U
PG2: G independent of U
PG3: G only associated with Y through T*

Fig 1. Top: Typical assumptions underlying the causal interpretation of an exposure *X* or treatment *T* on an outcome *Y* following standard regression analysis. Note that we assume NUC implies exchangeability: B = Bottom-left: Assumptions underlying a standard MR analysis using genetic variant G as an IV to estimate the causal effect of X on Y. C = Bottom-right: Assumptions underlying a standard pharmacogenetic analysis.

https://doi.org/10.1371/journal.pgen.1009783.g001

analysis if complete information on a patient's clinical state were available, this is seldom the case. A well known example (albeit in a non-pharmacogenetic context) where confounder adjustment failed is hormone replacement therapy, which was linked to increased cancer risk in observational data but not in subsequent randomized trials [6]. Thankfully, the need for adjustment can be circumvented if the purpose of the analysis is instead to compare the relative effectiveness of a treatment across genetic groups (see Fig 1). In the case of Clopidogrel and *CYP2C19*, typically one might assume that *CYPC219* variants *G*: do not predict whether an individual receives Clopidogrel *T*; are not associated with any confounders predicting Clopidogrel use and stroke, *Y*; and only affects stroke risk through their interaction with Clopidogrel. We will refer to these as the 'Pharmacogenetic' (PG) assumptions as a counterpart to the IV assumptions utilised by MR. A key difference exists between the role of the gene in MR and the role of the gene in pharmacogenetics: In MR, genes are assumed to directly influence a modifiable exposure. In Pharmacogenetics we can think of treatment as the exposure, and the genes are hypothesised to alter treatment *once* it is taken. In Fig 1 we denote the genetically altered treatment with the symbol T^*.

In recent work, Pilling et al [7] use data from GP-linked UK Biobank participants on Clopidogrel treatment to estimate its effect in different *CYP2C19* genetic subgroups and from this the number of strokes that could potentially be avoided if all individuals could experience the same benefit as the group with the most favourable genotype (through either dose modification or through switching to an alternative drug). In this paper we review the methodological underpinnings of the general PG approach, which utilises only individuals who are treated and relies on fairly strong baseline PG assumptions to estimate what we refer to as the 'Genetically Moderated Treatment Effect' (GMTE). When the PG assumptions are violated, we show that a robust but less efficient estimate of the GMTE that incorporates information on the population of untreated individuals can instead be used. In cases of partial violation, we clarify when Mendelian randomization and traditional confounder adjustment can also yield consistent estimates for the GMTE. A decision framework is then described to decide when a particular estimation strategy is most appropriate and how specific estimators can be combined to further improve efficiency.

Triangulation of evidence from different data sources, each with their inherent biases and limitations, is becoming a well established principle for strengthening causal analysis [8]. We call this framework '**T**riangulation **WI**thin a **ST**udy' (TWIST)' in order to emphasise that an analysis in this spirit is also possible within a single data set, using causal estimates that are approximately uncorrelated and reliant on different sets of assumptions. This makes their estimates easy to quantitatively combine if sufficiently similar to improve the precision and robustness of any findings. More broadly, it enables estimates to be qualitatively compared and contrasted, with expert judgement used to assess whether their assumptions are likely to have been met, in order to come to an overall conclusion about the totality of evidence. We illustrate these approaches by re-analysing primary-care-linked UK Biobank data relating to *CYP2C19* genetic variants, Clopidogrel use and stroke risk, and data relating to *APOE* genetic variants, statin use and Coronary Artery Disease (CAD).

2 Methods

Suppose that we are interested in evaluating the maximal effectiveness of a treatment T on an outcome Y using observational data. For simplicity we will assume initially that T is a binary treatment indicator so that, if prescribed, it is taken in full, that Y is a continuous or binary outcome variable and we are interested in estimating the treatment effect as a mean or risk difference contrast. A simple but naive way of estimating this effect would be to compare outcomes across those who are treated and those who are untreated. Borrowing terminology from the clinical trials literature, we refer to this as the 'As treated' (AT) estimate (Fig 1 top right). The AT estimate may not directly address our research needs for two reasons. The first reason is that, although we may understand many of the factors which influence whether an eligible individual is prescribed treatment by their doctor, there may be unmeasured variables, U, which influence both the decision to prescribe treatment and the outcome. Indeed, even if the treatment is truly effective in reducing the severity or risk of Y, it is highly likely that the population of treated individuals may still experience worse outcomes than those who are untreated. This would mean that the sign of the AT estimate could be positive, and thus qualitatively different than the true causal effect. This is classic confounding by indication. The second reason is that a pharmacogenetic investigation may suggest that the treatment does not in fact work for a certain proportion of the population at all, has a markedly reduced effectiveness, or increases the risk of side-effects.

We would like to estimate the difference in patient outcomes if all patients who take treatment experienced the 'full' effect, as experienced by those with the treatment-enabling genotype versus the reduced (or possibly zero) effect experienced by those with a treatment-inhibiting genotype. To realise such a benefit in practice, we could switch patients with the treatment-inhibiting genotype to an alternative medication which then works to the same extent as the 'full' effect of the original treatment. We will call this hypothetical quantity (or 'estimand') the Genetically Moderated Treatment Effect (GMTE).

2.1 The causal estimand and key identifying assumptions

To make the target of our analysis more explicit we will assume a simple model with a binary genotype G, where $G = 1$ denotes the treatment-enabling genotype and $G = 0$ denotes the treatment-inhibiting genotype. We now define a new treatment-moderating variable T^* which is equal to the product or interaction $T \times G$. We consider the following simple linear interaction model for the expectation (or mean) of outcome Y given treatment, T, genetic variant G, measured confounder Z and unmeasured confounder U:

$$E[Y|T,G,U] = \gamma_{Y0} + \beta_1 TG + \beta_0 T(1-G) + \gamma_{YG}G + \gamma_{YZ}Z + \gamma_{YU}U \tag{1}$$

$$= \gamma_{Y0} + \beta_0 T + (\beta_1 - \beta_0)T^* + \gamma_{YG}G + \gamma_{YZ}Z + \gamma_{YU}U \tag{2}$$

Under model (1), β_1 and β_0 reflect the treatment effect experienced by those with genotype $G = 1$ and $G = 0$ respectively and thus allows for genetically driven treatment effect heterogeneity. The parameter γ_{YG} represents the direct effect of G on Y and γ_{YU} represents the direct effect of U on Y. To clarify the causal estimand of interest we re-write model (1) as model (2). Using potential outcomes notation, we can express the GMTE estimand as the average causal effect if everyone could receive moderated treatment level $T^* = 1$ (i.e. the full or enhanced effect) versus if everyone could receive treatment level $T^* = 0$ (i.e. no enhanced effect):

$$\beta_{GMTE}(Y) = E[Y_i(T^* = 1) - Y_i(T^* = 0)] \tag{3}$$

This is equal to $\beta_1 - \beta_0$, the coefficient of T^* in model (2). We now define the key assumptions that will be leveraged by the various methods proposed in this paper. These assumptions are also represented by the causal diagram and corresponding association parameters in Fig 2:

- **Homogeneity (Hom)**: Individuals who take treatment with genotype level $G = 0$ experience no treatment effect all ($\beta_0 = 0$). Note that this is subtly different to Homogeneity assumption IV4 made in Mendelian randomization, which in this context would state that $\beta_1 = \beta_0$;
- **PG1**: An individual's genotype G is independent of the decision to take treatment, T, given all unmeasured confounders, U, of T and outcome Y ($\gamma_{TG} = 0$);
- **PG2**: An individual's genotype G is independent of confounders Z, U ($\gamma_{UG} = 0$);
- **PG3**: An individual's genotype G is independent of their outcome Y given treatment T and all unmeasured confounders U ($\gamma_{YG} = 0$);
- **No unmeasured confounding (NUC)**: All confounding variables U that predict T and Y have been measured and adjusted for (γ_{YU} or $\gamma_{TU} = 0$).

As previously stated, we will assume that the NUC assumption implies exchangeability between treatment groups, which rules out the presence of systematic selection bias or systematic loss to follow up in the data.

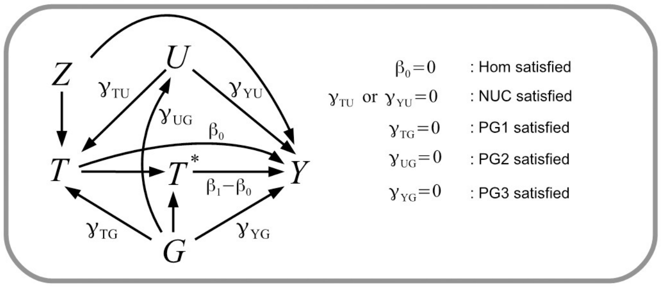

Fig 2. Causal diagram explaining the key assumptions leveraged by the methods proposed. The diagram and notation are consistent with outcome model (2) and the simulation simulation study. Here and throughout the paper, the variable Z represents measured confounders of T and Y and U represents unmeasured confounders.

https://doi.org/10.1371/journal.pgen.1009783.g002

2.2 Estimating the GMTE by correcting the As-Treated estimate

The As-Treated estimate suffers in general from confounding by indication, and a lack of specificity to the genetic variant driving the mechanistic interaction with the treatment. However, if the Hom and NUC assumptions are satisfied and either PG1 or PG3 are satisfied, then, the 'Corrected' As Treated estimate (CAT) can consistently estimate the GMTE. Put simply, if confounding bias can be addressed and the population treatment effect is driven entirely by the $G = 1$ subgroup, then the correct quantity can be estimated by scaling the treatment-outcome association by the proportion of treated individuals with the $G = 1$ genotype.

2.3 Estimating the GMTE in the treated population only

We next consider estimation of the GMTE in the general case where only assumptions PG1-PG3 hold. Together they imply that G is jointly independent of treatment and any unmeasured confounders, and they only affect the outcome through the treatment moderator variable T^*. Among the population of treated individuals, we can think of an individual's genotype as randomly allocating them to either moderated treatment level $T^* = 1$ or $T^* = 0$. This means we can calculate the GMTE using only treated individuals via the 'GMTE(1)' estimate. We use the additional subscript '(1)' in the estimate's nomenclature to denote that it conditions on $T = 1$.

2.4 A robust GMTE estimator

We next consider estimation of the GMTE under violations of PG1-PG3. Violation of PG1 implies that an individual's genotype directly influences the likelihood that they receive treatment. For example, it could be that those with a $G = 0$ genotype have an increased risk of side effects on treatment and choose to immediately come off the drug. An alternative explanation could be genetic population stratification [9]: e.g the allele frequency of the genetic variant G and the rate of treatment could simultaneously vary across individuals from different ethnic groups. An example of PG2 violation would be if the genetic variant increases the likelihood of an unmeasured risk factor for the outcome, and this risk factor also increases their likelihood of being treated. An example of PG3 violation would be if an individual's genotype directly affects the outcome through a pathway completely independent of either treatment or any confounding factor, which could be viewed as horizontal pleiotropy [10]. When any of these assumptions are violated the GMTE(1) estimate will reflect the genetically moderated effect of treatment plus the bias due to PG1–3 violation. Specifically, this bias would be the sum of:

- b_{PG1} via the $G \rightarrow T \leftarrow U \rightarrow Y$ pathway (due to violation of PG1);
- b_{PG2} via the $G \rightarrow U \rightarrow Y$ pathway (due to violation of PG2);
- b_{PG3} via the $G \rightarrow Y$ pathway (due to violation of PG3).

Whilst the bias contributions b_{PG2} and b_{PG3} are clear, bias contribution b_{PG1} is perhaps less so; it occurs because the GMTE(1) estimate explicitly conditions on treatment T, the presence of an association between G and T makes T a 'collider' [11]. This is one example of broader point: the RGMTE estimate is not robust to effect modification of any variable associated with T. Thankfully, when assumption PG1 is satisfied and no other effect modification is present, bias terms b_{PG2} and b_{PG3} can be consistently estimated and removed by incorporating information on the untreated population. This is achieved by calculating the equivalent GMTE(1) estimate for the untreated group (we call this the GMTE(0) estimate) and then subtracting this estimate from that in the treated group. We call this the 'Robust' genetically moderated treatment effect (RGMTE) estimate. Although assumption PG1 is key for the RGMTE estimate, we show that it can work if PG1 is violated but the NUC assumption is satisfied.

2.5 A 'Mendelian randomization' estimate

Given data on both treated and untreated individuals, it is possible to obtain an estimate for the GMTE by using the genetic variant G as an instrumental variable for the treatment moderator variable T^* directly, as in Mendelian randomization. In the context of a single gene, G, the MR estimate is the ratio of the gene-outcome association and the the gene-T^* association. The MR estimate is consistent for the GMTE if PG2-PG3 hold and either PG1 holds, or the Hom assumption holds.

In S1 Text we provide a formal justification of when the CAT, GMTE(1), RGMTE and MR estimates are consistent for the GMTE assuming outcome model (2).

2.6 Method summary and implementation

In Table 1 we: (i) give statistical formulae for the GMTE(1), GMTE(0), RGMTE, MR and CAT estimates; (ii) provide more detailed information on the sufficient assumptions each one relies upon to consistently estimate the GMTE (or in the case of the GMTE(0) estimate, zero); (iii) show how to test whether potential measured confounders of the treatment and outcome could bias each estimate; and (iv) give generic R psuedocode to obtain each estimate. To further clarify point (iii), take the GMTE(1) estimate as an example. In order to assess whether a potential confounder Z_1 could meaningfully bias its estimate, we calculate the GMTE(1) estimate but use Z_1 as the outcome in place of the true outcome Y. If this GMTE(1) estimate is significantly non-zero then it indicates a meaningful bias in the GMTE estimate with respect to the outcome, unless the confounder is adjusted for by treating Z_1 as an additional component of Z. This principle holds for all other GMTE estimates as well. Unlike the GMTE(1) and RGMTE estimators, the MR and CAT estimators both have a ratio form with the denominator dependent on G. For this reason they are more susceptible to bias and imprecision when the sample size is small and G has a low allele frequency.

When the outcome is continuous, the approaches can be implemented using linear regression to estimate the GMTE as a mean difference. With a binary outcome, we recommend

Table 1. Columns left to right show: Statistical formulae for GMTE(1), GMTE(0), RGMTE, MR and CAT estimates; Sufficient assumptions each one relies upon to consistently estimate the GMTE (or zero in the case of the GMTE(0) estimate); Estimate-specific confounder test statistics; generic R code to obtain each estimate. For the GMTE(0) estimate, $T_{CAT} = \hat{E}[G|T=1]T$, for the GMTE(0) estimate $T^- = 1 - T$, $T^{*-} = T^- G$, for the RGMTE estimate $T^* = TG$ and $\hat{T}^* = \hat{E}[T^*|G]$. Note that the GMTE(0) estimate does not directly target the GMTE, but rather zero under the PG assumptions.

Estimate	Statistical Formula	Sufficient Assumptions	Confounder Test	Fit in R
CAT				
$\hat{\beta}_{CAT}(Y)$	$\frac{\hat{E}[Y\|T=1]-\hat{E}[Y\|T=0]}{\hat{E}[G\|T=1]}$	$\{PG1 \cup PG3\} \cap NUC \cap Hom$	$\hat{\beta}_{CAT}(G) = 0, \hat{\beta}_{CAT}(Z) = 0$	Coef. of T_{CAT} in $Y \sim T_{CAT} + Z$
GMTE(0)				
$\hat{\beta}_{GMTE(0)}(Y)$	$\hat{E}[Y\|T=0, G=1] - \hat{E}[Y\|T=0, G=0]$	$PG3 \cap \{(PG1 \cap PG2) \cup NUC\}$	$\hat{\beta}_{GMTE(0)}(Z) = 0$	Coef. of T^{*-} in $Y \sim T^- + T^{*-} + Z$
GMTE(1)				
$\hat{\beta}_{GMTE(1)}(Y)$	$\hat{E}[Y\|T=1, G=1] - \hat{E}[Y\|T=1, G=0]$	$PG3 \cap \{(PG1 \cap PG2) \cup NUC\}$	$\hat{\beta}_{GMTE(1)}(Z) = 0$	Coef. of T^* in $Y \sim T + T^* + Z$
RGMTE				
$\hat{\beta}_{RGMTE}(Y)$	$\hat{\beta}_{GMTE(1)}(Y) - \hat{\beta}_{GMTE(0)}(Y)$	$PG1 \cup NUC$	$\hat{\beta}_{RGMTE}(Z) = 0$	Coef. of T^* in $Y \sim T + T^* + \hat{T}^* + Z$
MR				
$\hat{\beta}_{MR}(Y)$	$\frac{\hat{E}[Y\|G=1]-\hat{E}[Y\|G=0]}{\hat{E}[T^*\|G=1]-\hat{E}[T^*\|G=0]}$	$\{PG1 \cup Hom\} \cap \{PG2 \cup NUC\} \cap PG3$	$\hat{\beta}_{MR}(Z) = 0$	Coef. of $\hat{T}^*$ in $Y \sim \hat{T}^* + Z$

https://doi.org/10.1371/journal.pgen.1009783.t001

estimating risk differences directly using either a linear probability model, or using a logistic regression model to furnish estimates on the risk difference scale as an average marginal effect. With time-to-event data, we recommend analysing the data under an additive hazards model. Further details are provided S1 Text. We suggest to estimate mean difference, risk differences or additive hazard differences in order to obtain estimates for the GMTE from different estimators on the same scale, because these measures are collapsible. That is, they should remain constant when marginalised over unobserved confounders [12]. This is especially important for being able to effectively combine methods, as described in the next section.

2.7 Which estimates can be combined?

When two estimates are highly correlated, we gain little knowledge when they are observed to be similar. However, when two uncorrelated or weakly correlated estimates are similar, it gives credence to the hypothesis that they are estimating the same underlying quantity, and there is the potential to combine them into a single, more precise estimate. In S1 Text we show that the RGMTE and MR estimates are asymptotically uncorrelated. We also show that the CAT estimate is mutually uncorrelated with the GMTE(1) and RGMTE estimates, and uncorrelated with the MR estimate when G is independent of T. In cases where G and T are not perfectly independent, but G is a modest predictor of T (a highly plausible scenario in most pharmcogenetic contexts), the correlation between the MR and CAT estimate will be non-zero but practically negligible. The fact that most estimate-pairs are uncorrelated makes them easy to combine via a simple inverse variance weighted average or meta-analysis. In order to decide whether two uncorrelated estimates can be combined, we propose the use of a simple heterogeneity statistic. This procedure is illustrated in Fig 3 taking the GMTE(1) and CAT estimates as an example. Using each estimate we calculate their inverse variance weighted average and from this the heterogeneity statistic, $Q_{GMTE(1),\ CAT}$. If this statistic is less than the 1-α quantile of a χ^2_1 density (where α is the pre-specified significance threshold) then we judge the GMTE (1) and CAT estimates to be sufficiently similar to combine into a single estimate more efficient estimate. If $Q_{GMTE(1),CAT}$ is greater than 1-α threshold then the two estimates should be left separate. Along with single estimates, combined estimates that meet this heterogeneity statistic are colour coded blue (e.g. Fig 3 case (i)) and those which do not will be colour coded black (e.g. Fig 3 case (ii)). We stick to this convention for the remainder of the paper.

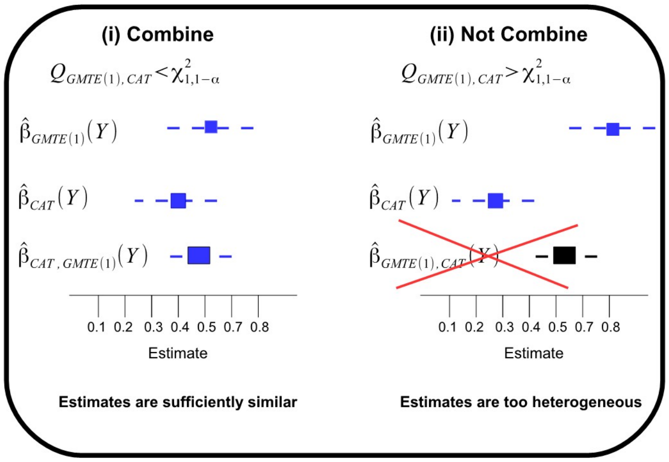

Fig 3. Two statistically uncorrelated estimates are homogenous enough to be meaningfully combined—case (i)—or are too heterogeneous to be combined—case (ii).

https://doi.org/10.1371/journal.pgen.1009783.g003

The RMGTE and MR estimates are in general highly correlated with the GMTE(1) estimate, and should therefore not be combined. S1 Text we show that the MR and RMGTE estimates can be viewed as complementary functions of the GMTE(1) and GMTE(0) estimates, and that the combined RGMTE/MR estimate is exactly equivalent to the GMTE(1) estimate when G is independent of T and the proportion of treated and untreated participants in the data is the same. Fig 4 shows a pictorial diagram of all single and combined estimates that can be derived using the above heterogeneity statistic criteria. This comprises four original estimates (CAT,GMTE1,RMGTE,MR), four 'paired estimates (CAT/GMTE1, CAT/RGMTE, CAT/MR, RGMTE/MR) and one 'triplet' estimate (CAT/RGMTE/MR), making nine in total. One possible analysis option would be to report all single and valid combined estimates which are sufficiently homogeneous according to a particular significance threshold. Another option would be to allow the GMTE(0) estimate to initially guide the analysis towards either the GMTE(1) estimate (and its possible combination with the CAT estimate) or the RMGTE estimate (and its possible combination with either MR estimate, the CAT estimates or both). Alternatively, some may be more comfortable with a qualitative assessment of the totality of evidence gleaned across the four distinct analysis procedures, using prior scientific knowledge to weigh up their individual importance after careful consideration given the plausibility of their key assumptions.

3 Simulation illustration

Trial data comprising a binary genotype G, treatment indicator T, observed covariate Z and a continuous outcome Y are simulated for n = 10,000 patients using the following data generating model which is consistent with the causal diagram in Fig 2:

$$\begin{aligned}
G_i &\sim Bern(p_G), \quad p_G = 0.3 \\
Z_i &\sim N(0,1) \\
U_i &\sim N(0,1) + \gamma_{UG} G_i \\
\eta_{Ti} &\sim \gamma_{T0} + \gamma_{TU} U_i + \gamma_{TG} G_i + \gamma_{TZ} Z_i + \epsilon_{Ti} \\
Pr(T_i = 1 | U_i, G_i) &= \text{expit}(\eta_{Ti}) \\
Y_i | T_i, G_i, Z_i, U_i &= \gamma_{Y0} + \beta_1 T_i G_i + \beta_0 T_i (1 - G_i) + \gamma_{YG} G_i + \gamma_{YZ} Z_i + \gamma_{YU} U_i + \epsilon_{Yi}
\end{aligned}$$

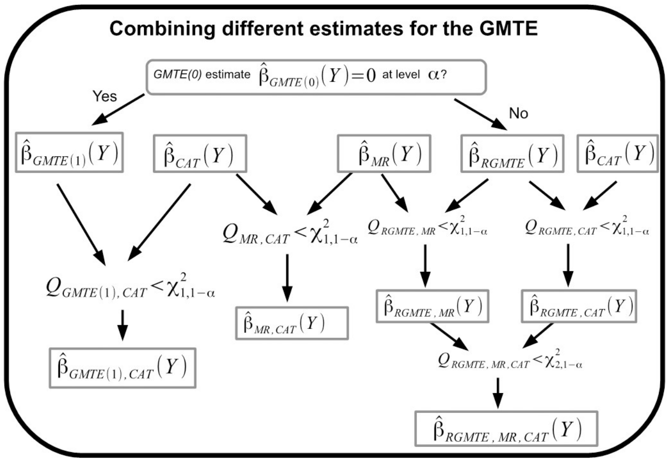
Combining different estimates for the GMTE
GMTE(0) estimate $\hat{\beta}_{GMTE(0)}(Y)=0$ at level α?
Yes
No
$\hat{\beta}_{GMTE(1)}(Y)$
$\hat{\beta}_{CAT}(Y)$
$\hat{\beta}_{MR}(Y)$
$\hat{\beta}_{RGMTE}(Y)$
$\hat{\beta}_{CAT}(Y)$
$Q_{MR,CAT}<\chi^2_{1,1-\alpha}$
$Q_{RGMTE,MR}<\chi^2_{1,1-\alpha}$
$Q_{RGMTE,CAT}<\chi^2_{1,1-\alpha}$
$Q_{GMTE(1),CAT}<\chi^2_{1,1-\alpha}$
$\hat{\beta}_{MR,CAT}(Y)$
$\hat{\beta}_{RGMTE,MR}(Y)$
$\hat{\beta}_{RGMTE,CAT}(Y)$
$\hat{\beta}_{GMTE(1),CAT}(Y)$
$Q_{RGMTE,MR,CAT}<\chi^2_{2,1-\alpha}$
$\hat{\beta}_{RGMTE,MR,CAT}(Y)$

Fig 4. A schematic diagram showing all possible 9 single, two-way or three-way combined estimators of the GMTE that can be calculated using the TWIST framework.

https://doi.org/10.1371/journal.pgen.1009783.g004

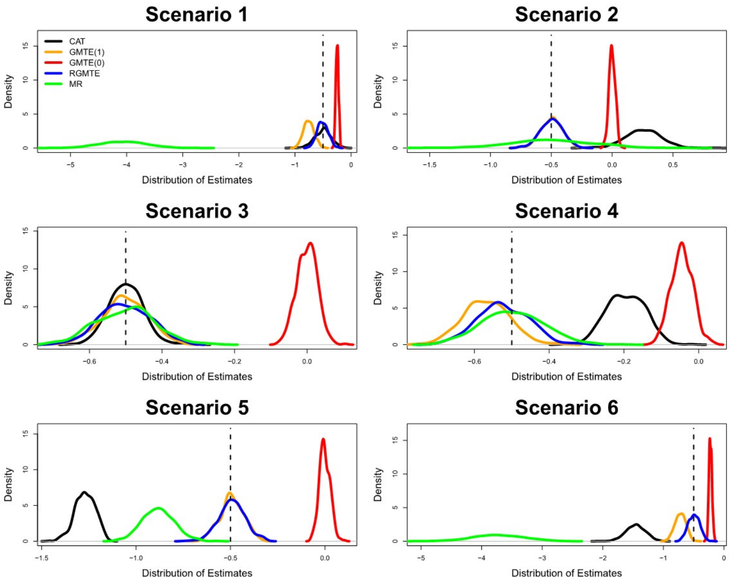

Fig 5. Distribution of estimates for the CAT, GMTE(1), GMTE(0), RGMTE and MR estimators across six simulation scenarios. In each case, the true GMTE is fixed at -0.5.

https://doi.org/10.1371/journal.pgen.1009783.g005

Under this model, assumptions PG1-PG3 are violated if γ_{TG}, γ_{UG} and γ_{YG} are non-zero respectively. The Hom assumption is violated when β_0 is non-zero. Finally the NUC assumption is violated if either γ_{TU} or γ_{YU} (or both) are non-zero. For simplicity we keep γ_{YU} fixed and non-zero and vary only γ_{TU}. Note that if the NUC assumption holds, then PG2 is in a sense automatically satisfied because U is no longer a confounder. However, in this case there may still be a path from G to Y via U. This would then form all or part of any PG3 violation.

Fig 5 shows the distribution of estimates for the GMTE obtained across 500 independent data sets and six simulation scenarios, using the CAT, GMTE(1), RGMTE and MR estimators. We also show the distribution of the GMTE(0) estimate in each case, as a helpful guide to understand the extent of bias that can be estimated from the data. In all scenarios the true GMTE is fixed at -0.5. Table 2 shows the mean point estimates, standard errors and 95% confidence interval coverage corresponding to the same six scenarios. For the five combined estimators, Table 3 shows: mean point estimates, mean standard errors, 95% confidence interval coverage and the proportion of times each combined estimator passes the heterogeneity test using a significance threshold of $\alpha = 0.05$.

In Scenario 1 of Fig 5 assumption PG3 is violated but all others (PG1, PG2, Hom, NUC) are satisfied. In this case both the CAT and RMGTE estimators are unbiased, with the RGMTE having the smallest standard error. In Table 3 we see that the combined RGMTE/CAT estimate is consequently unbiased with a standard error of 0.085, which is smaller than either the RMGTE or CAT estimates.

In Scenario 2 of Fig 5 the NUC assumption is violated but all others (PG1, PG2, PG3 Hom) are satisfied. In this case the GMTE(1), RGMTE and MR estimators are unbiased, with the GMTE(1) estimate being the most precise. In Table 3 we see that the combined RGMTE/MR estimate is consequently unbiased with a standard error near-identical to the GMTE(1) estimate, in line with the theoretical prediction outlined in S1 Text. In Scenario 3 of Fig 5 assumption PG1 is violated and (PG2, PG3, Hom, NUC) are satisfied. In this case

Table 2. Mean point estimates, standard errors and coverage (of 95% confidence interval) for the CAT, GMTE(1), GMTE(0) RGMTE and MR estimates across six simulation scenarios. In each case, the true GMTE is fixed at -0.5. Unbiased estimates and associated standard errors/coverages are highlighted in bold.

Scenario & Assumption(s) violated		**Estimator**				
		CAT	**GMTE(1)**	**GMTE(0)**	**RGMTE**	**MR**
1. PG3	Est.	**-0.502**	-0.755	-0.248	**-0.507**	-4.068
	S.E	**0.150**	0.093	0.025	**0.096**	0.350
	Cov^{ge}	**0.95**	0.22	0	**0.95**	0.00
2. NUC	Est.	0.255	**-0.498**	0.000	**-0.498**	**-0.497**
	S.E	0.140	**0.089**	0.025	**0.092**	**0.320**
	Cov^{ge}	0.00	**0.96**	0	**0.95**	**0.94**
3. PG1	Est.	**-0.499**	**-0.502**	0.000	**-0.502**	**-0.501**
	S.E	**0.050**	**0.063**	0.028	**0.069**	**0.081**
	Cov^{ge}	**0.95**	**0.95**	0	**0.95**	**0.96**
4. PG1 and NUC	Est.	-0.195	-0.568	-0.045	-0.523	**-0.497**
	S.E	0.050	0.061	0.028	0.067	**0.079**
	Cov^{ge}	0.00	0.79	0	0.93	**0.94**
5. PG2 & Hom	Est.	-1.270	**-0.493**	-0.001	**-0.492**	-0.883
	S.E	0.050	**0.063**	0.028	**0.069**	0.083
	Cov^{ge}	0.00	**0.95**	0	**0.94**	0.01
6. All except PG1	Est.	-1.458	-0.719	-0.223	**-0.496**	-3.728
	S.E	0.150	0.092	0.025	**0.095**	0.360
	Cov^{ge}	0.00	0.34	0	**0.96**	0.00

https://doi.org/10.1371/journal.pgen.1009783.t002

all estimators are unbiased. In Table 3 we show in this case that the most efficient unbiased estimate of all comes from combining the RGMTE, CAT and MR estimates. In Scenario 4 of Fig 5 PG1 and NUC are violated but the remaining assumptions (PG2, PG3, Hom) are satisfied. In this case only the MR estimate is unbiased. Consequently, no combined estimator is unbiased although the bias in the RGMTE/MR estimate is small. In Scenario 5 of Fig 5, PG2 and Hom are violated but (PG1, PG3, NUC) are satisfied. In this case the GMTE(1) and RGMTE estimators are unbiased, with the GMTE(1) estimator being the most efficient. No combined estimate is unbiased. In Scenario 6 of Fig 5 all assumptions except PG1 are violated. In this case only the RGMTE estimate is unbiased and, again, no combined estimate is unbiased.

In order to gauge the sensitivity of each estimator to the minor allele frequency of G, we repeat simulation Scenario 3 of Fig 5 for six values of p_G between 0.02 and 0.3. Fig 6 plots the mean standard error of the estimates in each case. We see clearly that the precision of all estimates is an increasing function of minor allele frequency. However, the loss in precision at low allele frequencies is strongest for the MR and CAT estimates. This is because they are both ratio estimates, with the denominator depending heavily on G.

In conclusion, our simulation study provides an empirical verification of the strengths and limitations of each approach, and when any two uncorrelated estimates can be effectively combined via a simple inverse variance weighted meta-analysis. Although the standard error of any estimate that combines the CAT and MR estimates requires G to be independent of T to be strictly valid (since this implies a zero correlation between their estimates) when this assumption is violated in Scenario 3 of Fig 5 it only induces a modest loss of coverage (e.g. 91% for the CAT/MR estimate and 92% for the CAT/RGMTE/MR estimate). Across the simulations the RGMTE is emerges as the most robust estimate.

Table 3. Mean point estimates, standard errors, coverage (of 95% confidence interval) and heterogeneity test rejection rates for the five combined estimates across six simulation scenarios. In each case, the true GMTE is fixed at -0.5. Unbiased estimates and associated standard errors/coverage are highlighted in bold.

Scenario & Assumption(s) violated		Estimator				
		RGMTE/MR	RGMTE/CAT	MR/CAT	GMTE1/CAT	RGMTE/MR/CAT
1. PG3	Est.	-0.754	**-0.506**	-1.041	-0.684	-0.684
	S.E	0.092	**0.080**	0.140	0.079	0.078
	Cov^{ge}	0.22	**0.96**	0.02	0.35	0.35
	$Q_p \geq 0.05$	0.00	0.94	0.00	0.70	0.00
2. NUC	Est.	**-0.498**	-0.274	0.134	-0.286	-0.286
	S.E	**0.089**	0.077	0.130	0.075	0.075
	Cov^{ge}	**0.96**	0.15	0.01	0.16	0.16
	$Q_p \geq 0.05$	0.94	0.00	0.42	0.00	0.01
3. PG1	Est.	**-0.501**	**-0.500**	**-0.500**	**-0.500**	**-0.500**
	S.E	**0.052**	**0.040**	**0.042**	**0.039**	**0.036**
	Cov^{ge}	**0.93**	**0.95**	**0.91**	**0.94**	**0.92**
	$Q_p \geq 0.05$	0.95	0.95	0.98	0.96	0.96
4. PG1 and NUC	Est.	0.512	-0.313	-0.282	-0.346	-0.350
	S.E	0.051	0.040	0.042	0.039	0.036
	Cov^{ge}	0.94	0.01	0.00	0.02	0.02
	$Q_p \geq 0.05$	0.94	0.04	0.07	0.00	0.01
5. PG2 & Hom	Est.	-0.652	-1.001	-1.167	-0.969	-0.979
	S.E	0.053	0.040	0.043	0.039	0.036
	Cov^{ge}	0.19	0.00	0.00	0.00	0.00
	$Q_p \geq 0.05$	0.05	0.00	0.01	0.00	0.00
6. All except PG1	Est.	-0.709	-0.761	-1.813	-0.913	-0.905
	S.E	0.092	0.081	0.140	0.079	0.079
	Cov^{ge}	0.37	0.11	0.00	0.00	0.00
	$Q_p \geq 0.05$	0.00	0.00	0.00	0.01	0.00

https://doi.org/10.1371/journal.pgen.1009783.t003

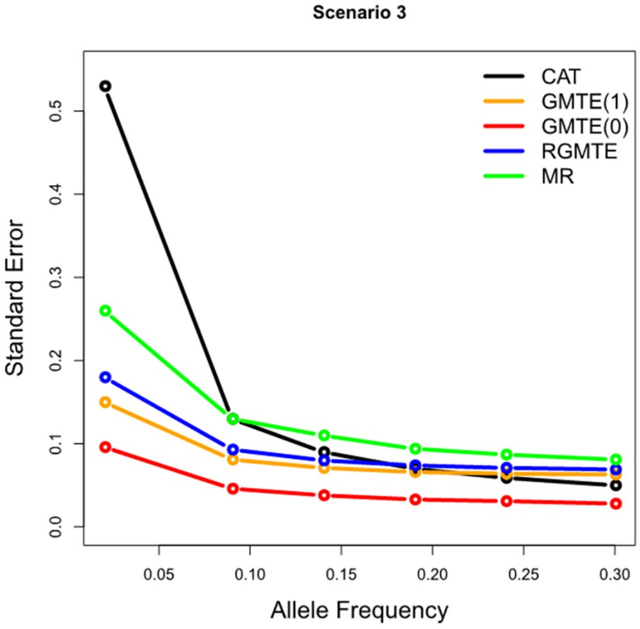

Fig 6. Mean standard error of the CAT, GMTE(1), GMTE(0), RGMTE and MR estimates for Scenario 3 as a function of the minor allele frequency of *G*.

https://doi.org/10.1371/journal.pgen.1009783.g006

4 Applied analyses

4.1 Clopidogrel, *CYPC219* & Stroke risk

Clopidogrel is a widely used anti-platelet therapy that impairs platelet aggregation with consequent reductions in risk of atherothrombotic events such as myocardial infarctions and ischemic strokes [13]. Clopidogrel is a pro-drug that requires activation by liver enzymes, primarily CYP2C19. Genetic variants in *CYP2C19* impair function with subsequently reduced Clopidogrel active plasma levels [14], and we have previously shown using primary care linked data on UK Biobank participants that carriers of these variants have increased risks of ischemic stroke and myocardial infarction (MI) whilst prescribed Clopidogrel [7]. In this work we calculated the population attributable fraction using established methods by analysing data on only those who were treated with Clopidogrel, but we revisit the analysis and apply the full TWIST decision framework proposed in this paper.

The UK Biobank study recruited 503,325 volunteers from the community who attended one of 22 assessment centres in England, Wales or Scotland between 2006 and 2010 [15]. Participants were aged 40 to 70 years at the time of assessment, and baseline assessment included extensive questionnaires on demographic, health, and lifestyle information. Blood samples were taken, allowing analysis of participant genetics. Ethical approval for the UK Biobank study was obtained from the North West Multi-Centre Research Ethics Committee. This research was conducted under UK Biobank application 14631 (PI: DM).

Linked electronic medical records from primary care are available for 230,096 (45.7%) of participants, which includes >57 million prescribing events between 1998 and 2017. Detailed description of the data extracted and limitations are available from UK Biobank. For this analysis we excluded 5,353 participants missing any genetics data, then 14,856 of non-European genetic ancestry, then 555 missing any *CYP2C19* loss of function genotype data, leaving 209,333 participants with sufficient primary care and genetic data. N = 198,868 never received a Clopidogrel prescription. N = 938 only ever received one prescription, so did not have sufficient exposure time for study. Of the 9,527 participants remaining, in 2,044 the prescribing frequency was less than once every 2 months, and these were also excluded. This left 7,483 participants with at least two Clopidogrel prescriptions for analysis. Baseline information on the included participants is shown in Table 4.

CYP2C19 loss-of-function (LoF) carriers (any *2-*8 alleles) had significantly increased ischemic stroke risk (Hazard Ratio (HR) 1.53: 95% CIs 1.04 to 2.26, p = 0.031) and separately MI (HR 1.14: 1.04 to 1.26, p = 0.008) whilst on Clopidogrel, compared to non-LoF carriers in Cox's proportional hazards regression models adjusted for age at first Clopidogrel prescription, sex, and the first 10 genetic principal components of ancestry. For this analysis non-LoF carriers constituted those with a 'normal' *CYPC219* genotype and those with the *CYP2C19* 17*

Table 4. Baseline data on UK Biobank participants in the Clopidogrel analysis set. *Based on hospital episode statistics data.

Variable	Prescribed Clopidogrel (n = 7,483)	Never Prescribed Clopidogrel n = 198,868
Age at recruitment	61.4 ± 6.2	56.5 ± 8.0
Age at first prescription	64.1 ± 7.3	-
Sex(Female%)	2,559(34.2%)	110,569(55.6%)
CYPC219 LoF carrier	2,145(28.7%)	56,043(28.2%)
Incident Ischemic Stroke diagnosis*	110 (1.5%)	2,078(1.0%)
Incident MI diagnosis*	1,822 (24.8%)	13,796(6.9%)

https://doi.org/10.1371/journal.pgen.1009783.t004

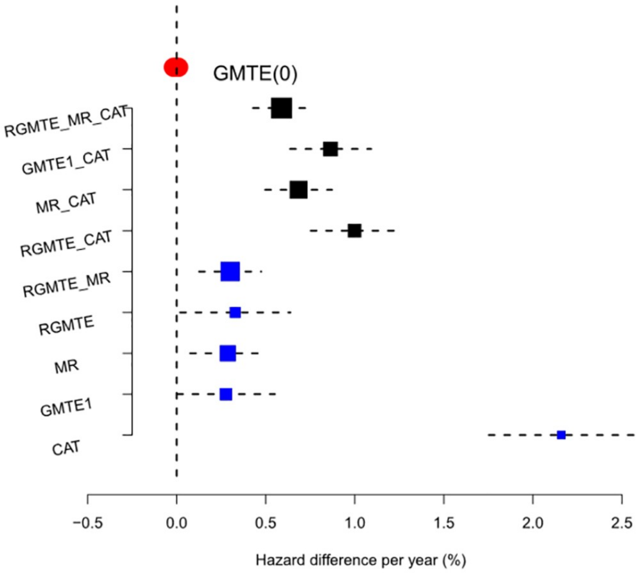

Fig 7. Forest plot of results for the Clopidogrel data. Blue squares show individual causal estimates as well as combined estimators that pass the heterogeneity test at the 5% level. Black squares show combined estimates that fail the heterogeneity test at the 5% level. Red bar shows the point estimate and confidence interval for the GMTE(0) estimate.

https://doi.org/10.1371/journal.pgen.1009783.g007

gain-of-function genotype. An in-depth analysis in our companion paper (Supplementary Table 3 in [7]) showed that normal and *^{*17}* individuals had a near-identical risk of stroke (HR = 0.99, p = 0.97) and that removing *^{*17}* individuals had little impact on the analysis estimates other than a loss in precision, since they constitute 22% of the population. For this reason we chose to keep the binary LoF/non-LoF genetic classification for the full TWIST analysis in the next section.

4.1.1 Estimating the GMTE. To estimate the GMTE in this case we modelled the time to stroke using an Aalen additive hazards model, as described in Section 2.4 and S1 Text. All models were adjusted for age at recruitment or first Clopidogrel prescription, sex, and the first 10 genetic principal components of ancestry. Fig 7 and Table 5 show the results for this analysis, which reflect the genetically moderated effect of Clopidogrel treatment on the hazard of stroke per year, expressed as a percentage. The GMTE(1) estimate suggests that being a

Table 5. Hazard difference estimates (LoF carriers versus non-carriers) on percentage scale for all single and combined estimates.

Estimator	Estimate (% scale)	S.E	p-value	Combined Estimates		
				Q statistic	Q p-value	Combine?
CAT	2.2	0.210	0.0e+00			
GMTE(0)	-3.9×10^{-3}	7.5×10^{-3}	6.1e-01			
GMTE(1)	0.28	0.140	4.8e-02			
MR	0.29	0.110	7.9e-03			
RGMTE	0.33	0.160	3.7e-02			
RGMTE/MR	0.3	0.089	7.5e-04	0.05	8.3e-01	Yes
RGMTE/CAT	1.0	0.130	2.0e-15	49	2.4e-12	No
MR/CAT	0.68	0.096	9.1e-13	64	1.3e-15	No
GMTE(1)/CAT	0.86	0.120	1.0e-13	56	5.9e-14	No
RGMTE/MR/CAT	0.59	0.082	6.6e-13	68	2.1e-15	No

https://doi.org/10.1371/journal.pgen.1009783.t005

CYP2C19 LoF carrier ($G = 1$) increases the risk of stroke by 0.28% (p = 0.048) compared to those without the LoF variant ($G = 0$). To put this figure in context, if we could reduce the LoF carrier's risk by this amount then, when multiplied by the 5264 LoF carrier patient years in the data, it would lead to an expected 13.2% reduction in the total number of strokes (or a reduction of 15 strokes from 110 to 95).

To test for potential bias in the GMTE(1) estimate, we calculate the GMTE(0) estimate in the untreated population. Thankfully, it is close to zero (Hazard diff = -0.0039%, p = 0.61), although slightly negative. Taken at face value, this suggest LoF carriers have a slightly reduced risk of stroke through pathways other than Clopidogrel use. Next we calculate the Corrected As Treated (CAT) estimate. As discussed, the validity of this method rests strongly on being able to identify all confounders of Clopidogrel use and stroke. With the data available, it was only possible to adjust for age, sex and genetic principal components and perhaps unsurprisingly, the CAT estimate is an order of magnitude larger (Hazard diff = 2.2%, $p \leq 2 \times 10^{-16}$). Consequently, the $Q_{CAT,GMTE(1)}$ statistic detects large heterogeneity and suggests that the CAT and GMTE(1) estimates should not be combined.

For completeness, we next calculate the RGMTE estimate for the GMTE hazard difference. Since this is itself the difference between the GMTE(1) and GMTE(0) estimates, and given they are of opposite sign, the RMGTE estimate is slightly larger at 0.33% (p = 0.037), suggesting 17 strokes could have been avoided. The MR estimate for the GMTE hazard difference is similar at 0.29% (p = 0.008). Heterogeneity analysis reveals that the MR and RGMTE estimates are sufficiently similar to combine into a more precise single estimate of the GMTE ($Q_{MR,RGMTE}$ = 0.8). The combined estimate is 0.3 ($p = 7.5 \times 10^{-04}$), or that 16 strokes could have been avoided. No other combination of estimates are sufficiently similar to combine.

4.2 Statins, *APOE* & CAD

We now apply our framework to estimate the extent to which genetic variation at the *APOE* locus modulates the risk of coronary artery disease (CAD) due to statin treatment using UK Biobank data. Our full data comprises 155,409 unrelated participants of European descent, with primary care data available (updated to March 2017) and up-to-date hospital admission data as of December 2020. Of this sample, we excluded: n = 11 participants with missing *APOE* genotypes; n = 6,456 non-regular statin users with less than four prescriptions per year or residuals from the linear regression for total statin prescriptions on years of statin treatment greater than 3 or less than -3; n = 1,273 non-statin users diagnosed with CAD at baseline (or prior to baseline); n = 4,566 participants starting statin after a doctor's diagnosis of coronary artery disease (CAD) based on the hospital admission records.

Among the included samples (n = 143,103), 57,682 (59.5%) were female. Of these, 46,179 (32.3%) were statin users, with a median of 9.4 (inter-quartile range: 6.6 to 13.5) statin prescriptions per year and a median of 5.6 (inter-quartile range: 1.2 to 9.9) years of statin treatment. Several SNPs were associated with LDL cholesterol response to statins based on a genome-wide association study, where the *APOE* e2 defining SNP rs7412 showed a larger LDL cholesterol lowering response to statins compared to e3e3s [16]. APOE genotypes (diplotypes essentially) were determined based on genotypes at rs7412 and rs429358. Inspecting the *APOE* genotype distribution, the majority of participants were classed as e_3e_3 (n = 83,813, 58.6%), followed by e_3e_4 (n = 33,597, 23.5%), e_2e_3s (n = 17,811, 12.4%), e_2e_4s (n = 3,616, 2.5%), e_4e_4s (n = 3,366, 2.4%), and e_2e_2s (n = 900, 0.6%). These groups are mutually exclusive. Summary statistics for statin users and non-statin users are presented in Table 6.

4.2.1 Results. Using the *e*3*e*3 group as a reference, we fitted Aalen additive hazard models within each mutually exclusive genetic group additionally adjusting for sex, age on statin or

Table 6. Baseline covariates, genetic data and incident CAD cases on statin users and non-users in UK Biobank.

	Statin Users (n = 46,179)	Non-Statin Users (n = 96,924)
Age starting statin (left)		
age at recruitment (right)	60.5± 7.9	54.7± 8.0
Sex (=Female)	20,921 (45.3%)	57,682 (59.5%)
APOE genotype		
e_3e_3	26,938 (58.3%)	56,875 (58.7%)
e_2e_2	258 (0.6%)	642 (0.7%)
e_2e_3	4,772 (10.3%)	13,039 (13.5%)
e_2e_4	1,065 (2.3%)	2,551 (2.6%)
e_3e_4	11,849 (25.7%)	21,748 (22.4%)
e_4e_4	1,297 (2.8%)	2,069 (2.1%)
Incident CAD (MI or angina) cases	7,259 (15.7%)	2,758 (2.8%)

https://doi.org/10.1371/journal.pgen.1009783.t006

age at recruitment, and the top 10 genetic principal components. For brevity, we focus on the results of the e2e3 versus e3e3 and e4e4 versus e3e3 analyses, which are shown in Table 7 and account for approximately 72% of the patient data. Estimates reflect the hazard or risk difference of a CAD event per year, expressed as a percentage. Only results of combined estimates that pass a heterogeneity test at the 5% level are shown. Equivalent estimates for the remaining genetic groups showed no evidence of a non-zero genetically moderated effect. Results for all genetic groups are given in Table 8.

4.2.2 e_4e_4 versus e_3e_3. Inspecting the e_4e_4 genetic subgroup first, the GMTE(1) estimate suggests that the risk of CAD could be reduced by 0.031% per year if e_3e_3 patients experienced the same treatment effect as e_4e_4 patients (p = 0.043). This estimate is valid if the e4e4 genotype only affects the risk of CAD through modulating the effectiveness of statins (i.e. assumptions PG1-PG3 hold). In order to probe this we calculate the equivalent GMTE(0) estimate in non-statin users. The e_4e_4 group is now seen to have a 0.011% larger risk of CAD than e_3e_3 (p = 0.07), which suggests that PG1-PG3 violation is possible. Furthermore, Table 6 shows clear differences in the allele frequencies between treatment groups. Since the RMGTE(0) and RGMTE(1) estimates are of opposite signs, the RGMTE estimate, which is robust to PG2-PG3 violation, infers the risk difference between e4e4 and e3e3's is larger at -0.037% per year

Table 7. Hazard difference estimates on the % scale for all single and valid combined estimates for the e2e3 and e4e4 genotype groups.

Estimator	Estimate	S.E	p-value	Expected Avoided CAD events in e3e3 group (95% CI)
		e4e4 versus e3e3		
CAT	3.9000	0.082	0.000	-
GMTE1	-0.0310	0.015	0.043	-85 (-168, -3)
MR	0.0069	0.017	0.690	19 (-76,115)
RGMTE	-0.0370	0.018	0.046	-103 (-204,-2)
GMTE0	0.0110	0.0063	0.073	-
RGMTE/MR	-0.0140	0.013	0.280	-39 (-108,31)
		e2e3 versus e3e3		
CAT	1.2000	0.0240	0.0e+00	-
GMTE1	-0.0100	0.0087	2.4e-01	-28 (-75,19)
MR	-0.0460	0.0110	3.4e-05	-128 (-189,-67)
RGMTE	-0.0055	0.0098	5.7e-01	-15 (-69,38)
GMTE0	-0.00014	0.0025	9.5e-01	-

https://doi.org/10.1371/journal.pgen.1009783.t007

Table 8. Hazard difference estimates on the % scale for all single and valid combined estimates.

Estimator	Estimate	S.E	p-value
	e2e2 versus e3e3		
CAT	18.800	0.407	0.000
GMTE1	0.004	0.036	0.906
MR	-0.006	0.047	0.892
RGMTE	-0.002	0.038	0.961
RGMTE/MR	-0.004	0.030	0.902
GMTE0	-0.003	0.010	0.725
	e2e3 versus e3e3		
CAT	1.180	0.024	0.000
GMTE1	-0.010	0.009	0.245
MR	-0.046	0.011	0.000
RGMTE	-0.006	0.010	0.571
GMTE0	0.000	0.002	0.954
	e2e4 versus e3e3		
CAT	4.700	0.100	0.000
GMTE1	-0.005	0.017	0.797
MR	-0.022	0.021	0.288
RGMTE	-0.001	0.020	0.977
RGMTE/MR	-0.011	0.015	0.452
GMTE0	0.002	0.005	0.666
	e3e4 versus e3e3		
CAT	0.580	0.011	0.000
GMTE1	-0.003	0.006	0.586
MR	0.011	0.007	0.107
RGMTE	-0.002	0.007	0.786
RGMTE/MR	0.005	0.005	0.346
GMTE0	0.001	0.002	0.615
	e4e4 versus e3e3		
CAT	3.860	0.082	0.000
GMTE1	-0.031	0.015	0.043
MR	0.007	0.017	0.694
RGMTE	-0.037	0.018	0.046
RGMTE/MR	-0.014	0.013	0.276
GMTE0	0.011	0.006	0.073

https://doi.org/10.1371/journal.pgen.1009783.t008

(p = 0.046). The MR estimate of the GMTE is also positive (0.0069%), but very close to zero (p = 0.69). This is, however, sufficiently similar to the RGMTE estimate at the 5% threshold for it to be combined with the MR estimate (despite being qualitatively different), and the combined value suggests a hazard difference of -0.014% per year (p = 0.28). The CAT estimate for the hazard difference in these data is a 3.9% increase per year. Its magnitude is so large compared to the other estimates that we could reasonably assume that adjustment for age, sex and genetic PCs is not sufficient to remove unmeasured confounding by indication. Consequently, no other estimate is sufficiently similar in order to combine with the CAT estimate, as shown in Fig 8. In the final column of Table 7 we translate the hazard difference estimate per year implied by the GMTE1, MR, RGMTE and combined RMGTE/MR estimate to give an expected number of CAD events that could be avoided if all 26,938 e3e3 statin user patients

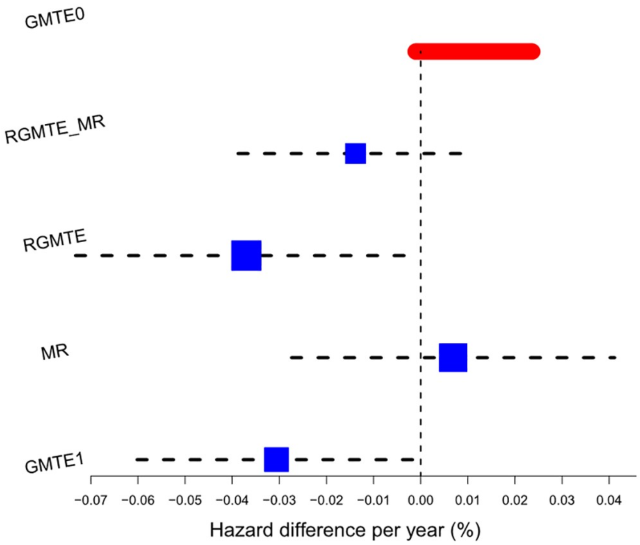

Fig 8. Hazard difference estimates for the e_4e_4 versus e_3e_3 analyses. Color coding the same as for Fig 7.

https://doi.org/10.1371/journal.pgen.1009783.g008

could receive the same benefit as the e4e4 patients, by multiplying the per-year risk reduction over the relevant 278,409 patients-years in the data. Using the RGMTE estimate for this risk reduction gives a figure of 103. The GMTE1 and combined RGMTE/MR estimates imply more modest reductions of 85 and 39 CAD events respectively.

4.2.3 e_2e_3 versus e_3e_3. Turning our attention to the e2e3 subgroup in Table 7 and Fig 9, we again see a large, non-credible CAT hazard difference estimate for the GMTE of a 1.2% per year between e2e3 and e3e3 groups. The GMTE(1), GMTE(0) and RGMTE estimates for the GMTE are all close to zero and non-significant at the 5% level. In contrast, the MR estimate for the GMTE suggests that e2e3's have a 0.046% reduced risk of CAD per year ($p = 3 \times 10^{-5}$). Using this estimate, the expected number of CAD events that could be avoided if all 26,938

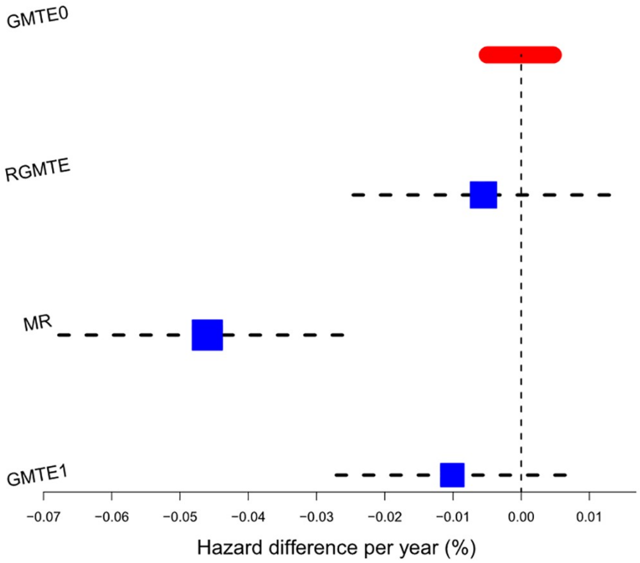

Fig 9. Hazard difference estimates for the e_2e_3 versus e_3e_3 analyses. Color coding the same as for Fig 7.

https://doi.org/10.1371/journal.pgen.1009783.g009

e3e3 patients could receive the same benefit as the e2e3 patients is 128. This is valid if we believe that assumptions PG2-PG3 hold, but either PG1 or the Hom assumption are violated. The GMTE1 and RGMTE estimates imply more modest reductions of 28 and 15 CAD events respectively. In this example, no two single uncorrelated estimates are sufficiently similar in order to combine.

5 Discussion

In this paper we propose the general TWIST framework for estimating the genetically moderated treatment effect that combines several distinct but complementary causal inference techniques. We propose a rudimentary decision framework for choosing when to combine approaches based on heterogeneity statistics. In practice, expert knowledge and prior evidence should also be leveraged to decide whether the particular assumptions of the causal estimation strategy are likely to be met, in order to put more or less weight on their findings. For example, if a variant is known to be associated with the outcome through another mechanistic pathway, then the PG3 assumption required for the GMTE(1) and MR estimates is likely violated, and the RGMTE estimate should be favored. Or, if it is known that those with the metabolically unfavourable genotype ($G = 0$) still benefit from treatment, then the homogeneity assumption is likely violated. This would then rule out the CAT estimate completely and one would need to be sure the PG1 assumption was satisfied when using the MR estimate.

In S1 Code we provide R code for fitting the TWIST framework for continuous, binary and time-to-event data as well as code used in the simulation study. Work is underway at https://github.com/lukepilling/twistR to produce a single R package to apply TWIST and visualise its results. Our inverse variance meta-analysis procedure for combining estimates is very simple, and simulations showed that it exhibited good statistical properties even when small correlations between constituent estimates were present. As future work, we plan to develop a more sophisticated procedure to explicitly account for this correlation within TWIST based on a Mahalanobis distance statistic, and to further develop the framework in several directions to address current limitations, some of which are now described.

5.1 Limitations and further work

We chose to illustrate the utility of the TWIST framework for combining similar estimates by demonstrating that it can increase precision. An alternative strategy would be to use multiple estimates to improve the robustness of any inference due to possible violations of variety of assumptions. For example, given a prior null hypothesis about the specific value of the GMTE, we would not reject the hypothesis it if was not rejected by any individual analysis. On the other hand, we could reject a proposed value of the estimand with increased confidence if it is rejected by multiple independent analyses that depend on assumptions that do not completely overlap. In future work we plan to develop a rigorous sequential testing procedure for TWIST that can control family wise error or false discovery rates. Since the majority of estimates reported within a TWIST analysis are statistically uncorrelated by design, multiplicity correction will be vital for this approach going forward. We thank reviewer 1 and 4 for these helpful suggestions.

The TWIST framework offers a means to combine statistically uncorrelated estimates that rely on overlapping sets of assumptions. If two estimates are similar enough to warrant combining into a single estimate, one hopes that this represents a more precise estimate of the true GMTE. However, there is always the possibility that both estimates are instead systematically biased in the same direction when there is a degree of overlap in their identifying assumptions and these assumptions are violated. In this case, combining them could give a more precise estimate of the wrong answer. Although we saw little evidence of this in simulation Scenarios 4–6 of Fig 5, further research is needed to understand the extent of this issue more clearly. We thank reviewer 4 for raising this important point.

In our analysis of the statin data, we estimated the GMTE in several mutually exclusive genetic groups, which resulted in an inevitable loss of precision. Efficiency could potentially be regained by collapsing genetic subsets together if they give similar estimates, or by making a linearity assumption about the magnitude of effect across genotypic groups (e.g. between e_3e_3, e_3e_4 and e_4e_4). This would not be defensible if the genetic groups were ordered with respect to the magnitude of their causal estimate, but would be defensible if genetic groups could be ordered by their effect on increasing drug metabolism. In the case of 3 genetic groups, G_i and T_i^* could take a value in {0,1,2}. This would enable the data to be pooled in order to target a combined estimand

$$\beta_{GMTE}(Y) = E[Y_i(T_i^*(m))] - E[Y_i(T_i^*(m-1))], \tag{4}$$

for all m in {1, 2}. If such a model were correct, it opens up the possibility of making the analysis even more robust to violations of the PG assumptions, because an additional causal parameter could be jointly estimated alongside the GMTE to reflect, for example the direct effect of G on Y. This is an important avenue for further research.

Although the CAT estimate can in principle consistently estimate the GMTE estimand, it relies heavily on the NUC assumption. In both applied analyses we were not able to sufficiently control for confounding by indication to deliver an estimate close to any other GMTE estimate, due to a lack of relevant covariate data. In future work we plan to revisit both analyses after collecting a much larger set of relevant information. More-sophisticated approaches such as Propensity Scores, matching methods and inverse probability weighting may then offer some utility [17]. So too may methods for multi-variable Mendelian randomization, where instead of directly adjusting for confounders of treatment and outcome, we instead adjust for their genetically predicted value. This latter approach could be more robust to collider bias [11].

The TWIST framework has parallels with the general theory of 'Evidence Factors' [18] for combining two or more observational associations estimates gleaned from the same data, which are susceptible to different biases. As far as we are aware, this approach has not been applied within the context of pharmacogenetics before, but a more detailed investigation of the connection between TWIST and Evidence Factors is an interesting topic for further research.

Supporting information

S1 Text. Document containing the important technical details on the TWIST framework, including consistency proofs for the linear case, and the implementation of TWIST with binary and time-to-event data.
(PDF)

S1 Code. Zip file containing code to implement the TWIST framework with continous, binary and time-to-event data.
(ZIP)

Author Contributions

Conceptualization: Jack Bowden, David Melzer.

Data curation: Luke C Pilling, Chia-Ling Kuo.

Formal analysis: Jack Bowden, Luke C Pilling, Deniz Türkmen, Chia-Ling Kuo.

Funding acquisition: David Melzer.

Investigation: Luke C Pilling, Chia-Ling Kuo.

Methodology: Jack Bowden.

Project administration: David Melzer.

Software: Jack Bowden, Luke C Pilling, Chia-Ling Kuo.

Supervision: Luke C Pilling, David Melzer.

Visualization: Jack Bowden, Chia-Ling Kuo.

Writing – original draft: Jack Bowden.

Writing – review & editing: Jack Bowden, Luke C Pilling, Deniz Türkmen, Chia-Ling Kuo, David Melzer.

References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 2003, 32(1):1–22.
2. Hernán MA. Beyond exchangeability: The other conditions for causal inference in medical research. *Statistical Methods in Medical Research* 2012, 21(1):3–5. https://doi.org/10.1177/0962280211398037 PMID: 22250015
3. Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP. CYP2C19 Genotype, Clopidogrel Metabolism, Platelet Function, and Cardiovascular Events: A Systematic Review and Meta-analysis. *JAMA* 2011, 306(24):2704–2714. https://doi.org/10.1001/jama.2011.1880 PMID: 22203539
4. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. *JAMA* 2016, 316(17):1818–1819. https://doi.org/10.1001/jama.2016.16435 PMID: 27802529
5. Veugelers PJ, Yip AM. Socioeconomic disparities in health care use: Does universal coverage reduce inequalities in health? *Journal of Epidemiology & Community Health* 2003, 57(6):424–428. https://doi.org/10.1136/jech.57.6.424 PMID: 12775787
6. Krieger N, Löwy I, Aronowitz R, Bigby J, Dickersin K, Garner E et al Hormone replacement therapy, cancer, controversies, and women's health: historical, epidemiological, biological, clinical, and advocacy perspectives. *Journal of Epidemiology & Community Health* 2005, 59(9):740–748. https://doi.org/10.1136/jech.2005.033316 PMID: 16100311
7. Pilling LC, Türkmen D, Fullalove H, Atkins JL, Delgado J, Kuo CL et al. Genetic variation in activating clopidogrel: longer-term outcomes in a large community cohort. *medRxiv* 2021.
8. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* 2017, 45(6):1866–1886.
9. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. *Current Protocols in Human Genetics* 2017, 95(1):1.22.1–1.22.23. https://doi.org/10.1002/cphg.48 PMID: 29044472
10. Hemani G, Bowden J, Davey Smith G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics* 2018, 27(R2):R195–R208. https://doi.org/10.1093/hmg/ddy163 PMID: 29771313
11. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology* 2017, 47(1):226–235.
12. Huitfeldt A, Stensrud MJ, Suzuki E. On the collapsibility of measures of effect in the counterfactual causal framework. *Emerging Themes in Epidemiology* 2019, 16(1):1. https://doi.org/10.1186/s12982-018-0083-9 PMID: 30627207
13. Mega JL, Close SL, Wiviott SD, Shen L, Hockett RD, Brandt JT et al. Cytochrome p-450 polymorphisms and response to clopidogrel. *New England Journal of Medicine* 2009, 360(4):354–362. https://doi.org/10.1056/NEJMoa0809171 PMID: 19106084

14. Simon T, Danchin N. Clinical impact of pharmacogenomics of clopidogrel in stroke. *Circulation* 2017, 135(1):34–37. https://doi.org/10.1161/CIRCULATIONAHA.116.025198 PMID: 28028061

15. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J et al Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* 2015, 12:1–10. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

16. Postmus I, Trompet S, Deshmukh HA, Barnes MR, Li X, Warren HR et al Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins *Nature Communications* 2014, 5(1):5068. https://doi.org/10.1038/ncomms6068 PMID: 25350695

17. Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* 2010, 25(1):1–21. https://doi.org/10.1214/09-STS313 PMID: 20871802

18. Rosenbaum P (2021) *Replication and Evidence Factors in Observational Studies*. New York: Chapman and Hall/CRC.